Запропоновано методику побудови математичних моделей для актуарних процесів та алгоритм оцінювання невідомих параметрів моделей із використанням байєсівського підходу. В якості математичного апарату використано узагальнені лінійні моделі, які представляють собою розширення лінійної регресії, коли розподіл випадкових величин відрізняється від нормального. На основі реальних статистичних даних та запропонованої методики, побудовано експериментальні моделі для прогнозування актуарних процесів

Ключові слова: узагальнені лінійні моделі, функція зв'язку, залишки, методи Монте-Карло, байєсівський підхід

Предложена методика построения математических моделей для актуарных процессов и алгоритм оценивания неизвестных параметров с использованием байесовского подхода. В качестве математического аппарата взяты обобщенные линейные модели, представляющие собой расширение линейной регрессии в случаях, когда распределение случайных величин отличается от нормального. На основании реальных статистических данных и предложенной методики построены экспериментальные модели для прогнозирования актуарных процессов

Ключевые слова: обобщенные линейные модели, функция связи, остатки, методы Монте-Карло, байесовский подход

УДК 519.766.4
DOI: 10.15587/1729-4061.2015.36486

МЕТОДИКА ПОБУДОВИ МАТЕМАТИЧНИХ МОДЕЛЕЙ АКТУАРНИХ ПРОЦЕСІВ

С. В. Трухан
Аспірант*
E-mail: svetlana.trukhan@gmail.com
П. І. Бідюк
Доктор технічних наук, професор*
E-mail: pbidyuke@gmail.com
*Інститут прикладного та системного аналізу
Національний технічний університет України
«Київський політехнічний інститут»
пр. Перемоги, 37, м. Київ, Україна, 03056

1. Вступ

Середовище актуарної діяльності – це сукупність процесів, пов'язаних з діяльністю страхових компаній, головною метою яких є фінансова компенсація наслідків випадкових подій, які несуть за собою матеріальні збитки. Однак, наприклад, азартні ігри та торгівля цінними паперами не є предметом розгляду у сфері страхування, оскільки у випадку з азартними іграми – учасники свідомо погоджуються на ризик, розуміючи, що ситуація може видатись вдалою або ж невдалою для їх матеріального положення. Аналогічна ситуація і з цінними паперами – учасник торгів може зазнати невдачі, але його ризики не страхуються, оскільки у такому випадку страхові компанії повинні розплачуватись за будь-яке невдале розміщення капіталу. Таким чином, у сфері страхування існує два основних види ризиків, перший – чистий ризик, а другий – спекулятивні. Предметом інтересу страховиків є тільки чисті ризики [1].

Страхова діяльність спрямована на перерозподіл грошових коштів та акумулювання їх безпосередньо для страхової діяльності, а з іншого боку – для інвестування цих коштів у різні галузі діяльності, що сприяє їх подальшому розвитку. Таким чином, виникають задачі аналізу та менеджменту фінансових ризиків і процесів інвестування з використанням сучасного апарату математичного моделювання, теорії оцінювання, прогнозування та ефективної підтримки прийняття рішень [2, 3].

Отже, в силу характеру та розмаїття діяльності страхових компаній ця сфера потребує розробки, удосконалення та впровадження у практику економіко-математичних моделей для обчислення, оцінювання і прогнозування в умовах невизначеності, ризику реалізації багатьох процесів, які зустрічаються у повсякденному житті фізичним особам та підприємствам різних форм власності і діяльності.

Робота посвячена розробці методики моделювання актуарних процесів з використанням узагальнених лінійних моделей (УЛМ) [4]. Також пропонується метод оцінювання параметрів УЛМ на основі байєсівського підходу. Наведено приклади застосування запропонованої методики побудови моделей з використанням фактичних даних.

2. Аналіз літературних даних та постановка проблеми

На практиці поширеними є методики побудови моделей типу авторегресії з ковзним середнім (АРКС), АРКС з ендогенними змінними (АРКСЕ) [5–7]. Однак, у представлених методиках поняття структури моделі не носить чіткого характеру, а визначенню нелінійності приділяється незначна увага, що може супроводжуватися отриманням хибних результатів при застосуванні до актуарних процесів. Основні передумови, на яких засновані математичні моделі прогнозування за допомогою ретроспективної вибірки за визначений період часу можна формулювати у вигляді таких трьох постулатів:

1. Ринок враховує все. Тобто значення показника є і наслідком, і вичерпним віддзеркаленням всіх рушійних сил ринку.

2. Рух цін підпорядкований деяким визначеним тенденціям. Життя ринку складається з періодів зростання та спадання цін, що чергуються, таким чином, щоб усередині кожного періоду відбувався розвиток пануючої тенденції, яка діє до тих пір, поки не почнеться рух ринку у зворотному напрямі.

3. Історія повторюється. «Ключ до розуміння майбутнього криється у вивченні минулого». Те, що певні конфігурації цін мають властивість з'являтися стійко та багаторазово, в різних часових періодах, є наслідком дії деяких стереотипів поведінки, властивих особі, яка приймає рішення, керує підприємством.

Виходячи із актуальності прогнозування та оцінювання актуарних процесів роботу присвячено дослідженню та розробці модифікованої методики побудови математичних моделей, виходячи із структури узагальнених лінійних моделей (УЛМ), які широко використовуються для аналізу страхових випадків, прогнозування продовження старих чи укладення нових страхових договорів, розробці тарифів та андерайтингу, цільовому маркетингу.

3. Цілі та задачі дослідження

Ставиться за мету виконання якісного аналізу існуючих структур математичних моделей для застосування у сфері страхування з метою їх подальшого використання для прогнозування розвитку досліджуваних процесів.

Для досягнення поставленої мети необхідно вирішити такі задачі:

– розробити ефективну методику побудови математичних моделей актуарних процесів у формі узагальнених лінійних моделей;

– запропонувати алгоритм оцінювання параметрів УЛМ для виконання подальших обчислювальних експериментів у сфері страхування актуарних ризиків;

– навести приклади застосування методики до побудови моделей актуарних процесів на основі фактичних даних.

4. Визначення структури математичної моделі

Розглянемо методику побудови математичних моделей для аналізу та прогнозування актуарних процесів у сфері страхування. Для побудови математичних моделей використовуємо узагальнені лінійні моделі, які утворюють досить узагальнений клас статистичних моделей, який включає лінійну та нелінійну регресію, дисперсійний та коваріаційний аналіз, Log-лінійні моделі для аналізу випадкових таблиць, нелінійні моделі типу пробіт/логіт, регресію Пуассона та деякі інші.

Основи методики побудови моделей часових рядів запропоновані Боксом і Дженкінсом у 1970-х роках і розвинуті в роботах [6, 7]. Модифікована методика побудови математичної моделі процесу:

1) системний аналіз процесу, для якого будується модель, на основі експертних оцінок протікання процесу, візуального дослідження вимірів вхідних і вихідних змінних, представлених часовими рядами, вивчення існуючих моделей та іншої доступної інформації;

2) попередня обробка експериментальних даних;

3) аналіз часових рядів на стаціонарність та можливу наявність нелінійностей за допомогою множини статистичних критеріїв якості;

4) визначити структури моделей-кандидатів, виходячи із структури УЛМ:

а) ідентифікація стохастичної, систематичної складових та функції зв'язку;

б) керуючись основними складовими УЛМ задати припущення стосовно випадковості, систематичності та вигляду функції зв'язку;

в) обчислити описові статистики та визначити величину відповідності даних вибраній моделі (класу моделей), використовуючи відповідні критерії (наприклад, критерій відношення правдоподібності);

г) визначити значимість предикторів за допомогою, наприклад, статистики Вальда та статистики міток;

д) оцінити характеристики інших елементів структури математичної моделі (наприклад, залишків);

5) вибрати метод (методи) оцінювання невідомих параметрів математичних моделей вибраних структур. Найчастіше це метод найменших квадратів (МНК), узагальнений метод найменших квадратів (УМНК), метод максимальної правдоподібності (ММП), метод Монте-Карло, байєсівский підхід;

6) вибрати кращу з оцінених моделей-кандидатів за допомогою множини статистичних критеріїв адекватності (якості) моделі.

Розглянемо докладніше кожний з наведених вище етапів з метою пояснення сутності та можливостей практичного застосування запропонованої методики.

4. 1. Системний аналіз процесу

У загальному випадку математична модель системи містить опис множини можливих станів останньої та закон переходу з одного стану до іншого. Але для того щоб побудувати таку математичну модель доцільно саме на етапі системного аналізу процесу виконати ідентифікацію об'єкта дослідження у вигляді системи, що перш за все передбачає визначення проблеми та проведення якісного аналізу процесу, а саме:

1) визначити сутність проблеми;

2) сформулювати можливі передумови та припущення;

3) визначити основні властивості об'єкта моделювання: структура, взаємозв'язки між його елементами;

4) при можливості сформулювати гіпотези, які відображають динаміку руху об'єкта та його взаємозв'язки із зовнішнім середовищем.

Аналіз процесу – це надзвичайно важливий етап, достовірне виконання якого потребує практичного досвіду дослідження реальних процесів різної природи та спрямовується на виконання таких задач [8–10]:

а) визначення кількості входів і виходів, тобто визначення розмірності моделі процесу;

б) встановлення логічних зв'язків між змінними та аналіз можливостей їх математичного опису (коректного об'єднання в одному математичному виразі);

в) визначення кількості зовнішніх збурень та їх типу (детерміноване чи стохастичне);

г) встановлення можливості розподілу процесу на окремі підпроцеси, які є простішими як з точку зору їх функціонування, так і з точки зору математичного опису; такий «умовний розподіл» в математиці найчастіше називають декомпозицією – це досить складний процес, який ґрунтується на спеціальних математичних методах;

д) якщо процес має ієрархічну структуру (верхній та нижній рівень функціонування), то необхідно чітко відокремлювати ці рівні, визначивши при цьому функції кожного з них і встановити які типи зв'язків існують між ними; наприклад, для технологічних процесів доречно будувати дерево ієрархії, що і розмежовує на два та більше рівнів функціонування і керування;

е) аналіз та використання знань із попередніх публікацій та досліджень щодо особливостей функціонування процесу, відомих законів та закономірностей його протікання, виявлення існуючих моделей процесу та досвіду його теоретичного чи експериментального дослідження;

ж) за наявності розроблених моделей досліджуваного процесу необхідно встановити їх недоліки та переваги, а також визначити можливість подальшого використання (модифікації); аналіз і використання існуючих моделей надає можливість суттєво скоротити час та інші витрати на побудову та використання моделі.

Ігнорування цього етапу призводить до ускладнень або до неможливості побудови моделі високого ступеня адекватності процесу, або до отримання хибної моделі, яка не відображає реальної сутності процесу дослідження та подальше використання отриманої моделі призведе, наприклад, до «неточного» прогнозу, а в подальшому – до корелювання незалежних змінних процесу. Отриману інформацію максимально використовують для попереднього оцінювання структури моделі або декількох моделей-кандидатів, параметри яких оцінюють за допомогою експериментальних даних. У процесі виконання аналізу функціонування досліджуваного процесу доцільно використовувати та порівнювати експертну інформацію з різних джерел. Це особливо стосується фінансово-економічних процесів, щодо яких може надходити інформація з протиріччями.

4. 2. Попередня обробка експериментальних даних

З практичної точки зору експериментальні дані – це випадкові величини або величини наближені до них, тому в ході експерименту існує проблема отримання інформації спотвореної певними перешкодами, результуючі дані можуть суттєво різнитись за одних і тих же умов. Мета попередньої обробки даних полягає у відсіюванні грубих похибок і оцінюванні достовірності експериментальних результатів, перевірка відповідності результатів вимірювання одному із законів множини експоненціальних розподілів.

Процес попередньої обробки експериментальних (статистичних) даних, як правило, включає такі операції:

а) нормування та візуальний аналіз даних і, за необхідності, їх корегування; нормування даних означає їх логарифмування або приведення до зручного діапазону їх зміни;

б) корегування даних полягає у заповненні пропусків та зменшенні викидів (екстремальних значень), що виходять за основний діапазон значень змінних;

в) формування перших або різниць вищих порядків, які необхідні для аналізу відповідних складових часового ряду.

Використання кінцевих різниць дає можливість будувати моделі для швидкості та прискорення основної змінної. Часто із значень ряду віднімають його середнє для того, щоб отримати можливість працювати з відхиленнями, а не повними значеннями змінних. Це необхідно робити, наприклад, при побудові моделей у просторі станів. Для оцінювання середнього існує метод оновлюваного середнього, який полягає у використанні рекурентної формули для обчислення величини середнього арифметичного. Якщо випадкова величина X надходить у вигляді дискретних вимірів та для $(N-1)$-го виміру обчислено середнє значення, то поява нового виміру змінює попереднє середнє значення на величину $\frac{1}{N}(X_N - X_{N-1})$. Таким чином, при згладжуванні за цим методом у кожній точці на числовій осі експериментальне значення заміняється на величину середнього, що розраховане на конкретний момент часу. Послідовність таких середніх значень є статистичним рядом без перешкод щодо змінної x, який використовується при подальшій обробці. Застосування того чи іншого методу підготовки даних для моделювання визначається в кожному випадку по-своєму.

4. 3. Аналіз наявності нелінійностей

Для розв'язання проблеми, пов'язаної з визначенням наявності нелінійностей у досліджуваному процесі, їх характеру використовують візуальний аналіз даних та формальні тести. За допомогою візуального аналізу виявляють існування ділянок з лінійним або нелінійним трендом, в якійсь мірі наявність гетероскедастичності та значних викидів, які можуть суттєво впливати на якість моделі [8].

Існує також ряд формальних тестів на наявність нелінійності. Розглянемо простий тест для визначення наявності нелінійності. Цей тест можна застосувати у випадку, коли до уваги можна взяти декілька вибірок спостережень для одного і того ж процесу:

$$\hat{F} = \frac{\frac{1}{m-2}\sum_{i=1}^{m} n_i (\bar{y}_i - \hat{y}_i)^2}{\frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2},$$

де $\bar{y}_i$ – середнє значення для $i-$ї групи (вибірки або групи) даних; $\hat{y}_i$ – середнє для лінійної апроксимації даних; $m-$ число груп даних; n_i- число вимірів в $i-$й групі; n – загальне число вимірів. Якщо статистика $\hat{F}$ з $v_1 = m-2$ та $v_2 = n-m$ степенями свободи досягає або перевищує рівень значимості, то гіпотезу щодо лінійності необхідно відхилити. Недоліком даного підходу є те, що для його застосування необхідно мати кілька (не менше трьох) груп даних для одного і того ж процесу, які можна отримати в результаті виконання повторних експериментів.

Наявність нелінійності можна встановити також за допомогою вибіркових нелінійних кореляційних функцій (НКФ), тобто кореляційних функцій, розрахованих за вибірками експериментальних (статистичних) даних. Наприклад, якщо дискретна НКФ

$$r_{yx^2}(s)=r_{y(k)x^2(k-s)}=\frac{1}{N}\frac{\sum_{k=s+1}^{N}\left\{\left[y(k)-\overline{y}\right]\left[x(k-s)-\overline{x}\right]^2\right\}}{\sigma_y\sigma_x^2},$$

$$s=0,1,2,3,\ldots,$$

містить значення, які суттєво відрізняються від нуля в статистичному смислі, то процес містить квадратичну нелінійність стосовно регресора x.

Наявність нелінійного детермінованого тренду у процесі можна визначити шляхом оцінювання рівняння:

$$y(k)=a_0+c_1k+c_2k^2+\ldots+c_mk^m,$$

яке представляє собою поліном порядку m стосовно часу. Якщо хоча б один із коефіцієнтів c_i, $i=1,\ldots,m$ є статистично значимим, то гіпотеза щодо відсутності тренду відхиляється. Якщо тренд відносно швидко змінює свій напрям руху і для нього трудно знайти адекватний функціональний опис, то застосовують моделі випадкових трендів, які ґрунтуються на комбінаціях випадкових величин [7].

Автоматично оцінює структуру математичної моделі *метод групового врахування аргументів* (МГВА), який багаторазово застосовано до широкого класу процесів; його успішно застосовують і сьогодні до моделювання процесів різної природи з нелінійностями та нестаціонарностями. Подальшим розвитком даного методу є нечіткий МГВА, який ґрунтується на нечіткому представленні параметрів оцінюваної моделі [8].

4. 4. Формування інших елементів структури моделі

На даному етапі необхідно вибрати структури моделей кандидатів. Поняття структури математичної моделі розглядається, виходячи із складових УЛМ та включає наступне:

а) визначення *стохастичної складової* – залежної змінної, яка розподілена згідно одного з законів розподілу із сімейства експоненціальних законів розподілу з середнім μ. Цю компоненту називають ще розподілом вихідної змінної.

б) встановлення *систематичної складової* – p-незалежних змінних, які об'єднуються в один «лінійний предиктор» [7]:

$$\eta=X\cdot\beta.$$

в) визначення характеру *функції зв'язку*, виходячи із розглянутої їх класифікації. Вона відображає взаємозалежність між припущенням випадковості та систематичності;

Крім цього, до структури моделей належать показники, які визначають:

а) порядок моделі (найвищий порядок рівнянь, що його утворюють);

б) розмірність (кількість рівнянь моделі);

в) залишки в УЛМ, які використовуються для дослідження адекватності моделі прогнозування, вибору функції зв'язку, функції дисперсії та елементів лінійного предиктора;

г) можливі нелінійності та їх тип;

д) зовнішні збурення та їх тип (детерміновані або випадкові; адитивні та мультиплікативні).

Коефіцієнт кореляції, а в загальному випадку кореляційна функція, дає можливість встановити факт існування зв'язку між змінними. Кореляційна матриця дає можливість встановити існування зв'язку між залежною (ендогенною) змінною та незалежними (екзогенними) змінними у правій частині. Саме тому, для того щоб визначити питання включення незалежних змінних (регресорів) в праву частину рівняння, обчислюють коефіцієнт кореляції між залежною та відповідними незалежними змінними.

Вважається, що сукупний вплив не вимірюваних випадкових факторів можна описати, в деякій мірі, за допомогою випадкової змінної $\varepsilon(k)$. Оскільки вона не вимірюється, то оцінити її значення (похибку моделі або залишок) можна тільки після оцінювання коефіцієнтів моделі, тобто

$$\hat{\varepsilon}(k)=e(k)=y(k)-\hat{y}(k),$$

де $\hat{y}(k)-$ оцінка змінної $y(k)$, отримана за допомогою моделі; $y(k)-$ фактичний вимір.

Крім кореляційної матриці для аналізу змінних моделі використовують описові статистики, наприклад, такі: коефіцієнт асиметрії, коефіцієнт ексцесу, показник статистики Жак-Бера, що стосується перевірки гіпотези про нормальний розподіл.

Таким чином, цей етап закінчується формуванням структур кількох моделей-кандидатів з різними законами розподілу: нормальний, гамма, розподіл Пуассона та відповідним видом функції зв'язку (логарифмічна, тотожна). Кандидатів може бути кілька, оскільки встановити структуру за один раз, як правило, неможливо. Загалом побудова моделі високого ступеня адекватності – це ітераційний процес, який вимагає значних зусиль. На наступному етапі оцінюють параметри моделей-кандидатів.

4. 5. Оцінювання коефіцієнтів моделей-кандидатів

На *четвертому етапі* оцінюють коефіцієнти (параметри) рівняння, використовуючи принцип економії або збереження. Цей принцип означає, що *кількість коефіцієнтів, що оцінюються, не повинна перевищувати їх необхідне число* ("необхідність" можна визначити, наприклад, як необхідність збереження в моделі основних статистичних характеристик процесу) [9, 10].

Оцінювання параметрів узагальненої лінійної моделі зазвичай виконується за допомогою методу найменших квадратів. Для нормально розподілених моделей це еквівалентно оцінюванню за методом максимальної правдоподібності. Але, при моделюванні процесів будь-якої природи необхідно пам'ятати, що поведінку процесу необхідно апроксимувати за допомогою рівнянь, а не старатися описати його до найменших дрібниць.

В процедурі оцінювання часто використовують не абсолютні значення змінних, а їх відхилення від середнього, тобто

$$y(k) = Y(k) - \mu_y,$$

де $Y(k)$ – значення виміру, μ_y – середнє значення ряду. Якщо для оцінювання параметрів використовується рекурсивна процедура, то поточне середнє можна обчислювати за формулою:

$$\mu_y(k) = \mu_y(k-1) + \frac{1}{k}[y(k) - \mu_y(k-1)].$$

Найбільш поширеними методами оцінювання параметрів моделі є такі: метод найменших квадратів (МНК); зважений метод найменших квадратів (ЗМНК); метод максимальної правдоподібності (ММП); методи Монте-Карло (ітеративні та неітеративні); метод моментів. Так, оцінки МНК обчислюють за допомогою виразу:

$$\hat{\theta} = \left[\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y},$$

де $\theta[p]$ – вектор оцінок параметрів вимірності p; $\mathbf{X}[N \times p]$ – матриця вимірів; $\mathbf{y}[N]$ – вектор вимірів залежної змінної. В квадратних дужках вказана розмірність векторів і матриці. Елементи матриці вимірів обчислюються по-своєму для кожної конкретної моделі.

Оскільки, застосування МНК в деяких випадках призводить до отримання зміщених оцінок через чутливість оцінок до різких викидів, які зустрічаються у вихідних даних. Тому альтернативою для МНК є метод максимальної правдоподібності – один із основних методів, який застосовується для оцінювання параметрів УЛМ. Для нормальної похибки, логарифмічну функцію правдоподібності l, виходячи з n-спостережень, можна представити у вигляді:

$$-2l = n\log(2\pi\sigma^2) + \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{\sigma^2}.$$

Для фіксованого σ^2 максимізація параметра l еквівалентна мінімізації суми квадратів $\sum(y-\mu)^2$ для відхилень μ; так, для лінійної моделі:

$$\eta_i = \mu_i = \sum_{j=1}^{p} x_{ij}\beta_j.$$

В останні десятиліття набув популярності метод Монте-Карло для оцінювання невідомих параметрів класу УЛМ. Основна ідея методу Монте-Карло полягає у рівномірному розбитті інтервалу спостереження даних $[t_i; t_{i+1}]$ на множину менших відрізків шляхом введення неспостережуваних (згенерованих) величин у проміжні моменти часу [11].

Алгоритм оцінювання параметрів моделей функціонує таким чином:

– генерування значення X_u^{i+1} з розподілу $P(X_u | X_0, \theta^i)$, де X_0 – експериментальні значення процесу;

– генерування значення параметрів θ^{i+1} з розподілу $P(\theta | X_0, X_u^i)$.

При досить слабких умовах регулярності згенерований таким чином статистичний ряд має граничний стаціонарний розподіл $P(X_u, \theta | X_0)$, який за допомогою тривіальних арифметичних операцій перетворюється в $P(\theta | X_0)$. Недолік методів Монте-Карло – велика ресурсоємність. Проте методи цієї групи широко використовуються завдяки універсальності, хорошій масштабованості, здатності ураховувати неспостережувані змінні, незначним похибкам оцінювання, а також можливості застосування паралельних обчислень для прискорення процесу оцінювання. Зазначимо, що перевірити виконання наведених умов можна тільки після оцінювання коефіцієнтів моделі, а до оцінювання можна тільки постулювати їх виконання. Тобто після оцінювання моделі оцінка значень випадкового процесу визначається похибками моделі: $\hat{\varepsilon}(k) = e(k) = y(k) - \hat{y}(k)$, що дає можливість виконати аналіз характеристик випадкового процесу $\{\varepsilon(k)\}$.

4. 6. Діагностика моделей – вибір кращої з множини оцінених кандидатів

На п'ятому етапі аналізується якість моделі, тобто виконується перевірка оцінених кандидатів на адекватність процесу. Діагностика складається з наступних кроків:

а) *Візуальне дослідження графіка похибок* моделі $e(k) = y(k) - \hat{y}(k)$, де $\hat{y}(k)$ – оцінка змінної, отримана за допомогою побудованої моделі. На графіку не повинно бути значних викидів та довгих інтервалів, на яких похибка приймає великі значення (тобто довгих інтервалів суттєвої неадекватності). Наприклад, у випадку побудови УЛМ, коли закон розподілу залежної змінної було обрано випадково, без аргументації та візуального аналізу, то в результаті використання такої моделі буде отримано значні відхилення від реальних даних при прогнозуванні, що підтверджується критичним значенням похибки і хибною оцінкою параметрів [7].

б) *Похибки моделі не повинні бути корельовані* між собою. Для аналізу наявності кореляції між значеннями похибок необхідно обчислити АКФ та ЧАКФ для ряду $\{e(k)\}$ і за допомогою Q – статистики визначити ступінь корельованості (наприклад, Q – статистика вважається несуттєвою до рівня 10 %).

Крім того, корельованість похибок визначають за допомогою статистики Дарбіна-Уотсона (DW), яка розраховується за формулою:

$$DW = 2 - 2\rho,$$

де $\rho = E[e(k)e(k-1)]/\sigma_e^2$ – коефіцієнт кореляції між сусідніми значеннями похибки; σ_e^2 – дисперсія послідовності похибок $\{e(k)\}$. Таким чином, при повній відсутності кореляції між похибками $DW = 2$ – це ідеальне значення. Граничними значеннями для DW є 0 (при $\rho = 1$) та + 4 (при $\rho = -1$).

в) Перевірка значимості параметрів моделі. *Статистика Стьюдента або* t – *статистика* (випадкова величина, що має t – розподіл), яка використовується для визначення значимості оцінки кожного коефіцієнта в статистичному сенсі, визначається за виразом:

$$t = \frac{\hat{a} - a^0}{SE_{\hat{a}}},$$

де $\hat{a}$ – оцінка коефіцієнта моделі; a^0 – нуль-гіпотеза (початкова гіпотеза) щодо цієї оцінки; $SE_{\hat{a}}$ – стандартна похибка оцінки. За нуль-гіпотезу щодо значимості оцінки можна висувати будь-яку: що коефіцієнт значимий, тобто $H_0 : a^0 \neq 0$, або не значимий: $H_0 : a^0 = 0$.

Статистична теорія перевірки гіпотез пропонує висувати нуль-гіпотезу, яка є протилежною бажаному результату. У даному випадку бажаним результатом є значимість коефіцієнтів математичної моделі. Таким чином, необхідно висувати нульову гіпотезу, що коефіцієнт не значимий. Це дає можливість коректно підійти до визначення значимості оцінок коефіцієнтів та дещо спростити розрахунки.

г) *Коефіцієнт множинної детермінації* R^2, який обчислюється так:

$$R^2 = \frac{\mathrm{var}(\hat{y})}{\mathrm{var}(y)} = 1 - \frac{SSE}{SST},$$

де $\mathrm{var}(\hat{y})$ – дисперсія залежної змінної, оціненої за допомогою побудованої моделі; $\mathrm{var}(y)$ – дисперсія вимірів залежної змінної; $SSE = \sum_{k=1}^{N} [y(k) - \hat{y}(k)]^2$ – сума квадратів похибок (залишків) моделі (sum of squared errors); $SST = \sum_{k=1}^{N} [y(k) - \bar{y}]^2$ – загальна сума квадратів (total sum of squares); $\bar{y}$ – середнє значення; $SST = SSE + SSR$, де $SSR = \sum_{k=1}^{N} [\hat{y}(k) - \bar{y}]^2$ – загальна сума квадратів для регресії. Очевидно, що найкращим значенням є $R^2 = 1$, тобто, коли дисперсії вимірів змінної, та цієї ж змінної, оціненої за рівнянням, збігаються. Цей параметр можна трактувати, також, як *міру інформативності моделі*, якщо вибрати за міру інформативності дисперсію. Таким чином, R^2 показує рівень інформативності моделі по відношенню до інформативності вибірки даних, за допомогою якої вона була оцінена.

е) *Сума квадратів похибок* для вибраної моделі повинна бути мінімальною, тобто,

$$\sum_{k=1}^{N} e^2(k) = \sum_{k=1}^{N} [\hat{y}(k) - y(k)]^2 \to \min_{\hat{\theta}}$$

порівняно з усіма іншими моделями.

є) Для оцінювання адекватності моделі також використовують *інформаційний критерій Акайке:*

$$AIC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + 2n$$

та критерій Байєса-Шварца

$$BSC = N \ln\left(\sum_{k=1}^{N} e^2(k)\right) + n \ln(N),$$

де $n = p + q + 1$ – кількість параметрів моделі, які оцінюються за допомогою статистичних даних (p – кількість параметрів авторегресійної частини моделі; q – число параметрів ковзного середнього; 1 з'являється тоді, коли оцінюється зміщення (або перетин, тобто a_0).

У правій частині критеріїв Акайке і Байєса-Шварца міститься сума квадратів похибок, а тому за цими критеріями вибирають ту модель, для якої критерії приймають найменші значення. Введення нового регресора приводить до збільшення критерію (при цьому збільшується n), але, разом з тим, зменшується сума квадратів похибок і критерій в цілому зменшується. Якщо регресор не покращує модель, то критерій збільшується. Необхідно також зазначити, що асимптотичні властивості для довгих виборок кращі у критерія Байєса-Шварца, тобто, його рекомендують застосовувати при відносно великих значеннях N ($N > 100$).

ж) Коефіцієнт Дарбина-Уотсона відображає адекватність побудованої моделі та обчислюється за формулою

$$DW = \frac{\sum_{t=2}^{N} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{N} \varepsilon_t^2},$$

де ε – вектор залишків (різниця між значеннями отриманими за моделлю та фактичними), при цьому $DW \in [0;4]$. Для кращої моделі $DW \to 2$; це означає, що залишки моделі між собою не автокорелюють.

з) Окрім згаданих параметрів, для визначення адекватності моделі в цілому використовують F – *статистику Фішера*, яка пропорціональна відношенню:

$$F \sim \frac{R^2}{1 - R^2}.$$

Якщо $R^2 \to 1$, то $F \to \infty$. Порядок застосування F – статистики такий же, як і t – статистики. Нуль-гіпотезою є в даному випадку припущення про те, що модель неадекватна в цілому, тобто,

$$H_0 : a_1 = a_2 = ... = a_p = 0$$

проти альтернативної гіпотези:

H_1: хоча б одне значення a_i відмінне від нуля в статистичному смислі.

Значення $F_{крит}$ знаходять із таблиць для F – розподілу.

Крім цього, часто для формування статистичного висновку використовують байєсівський коефіцієнт, який представляє собою відношення апостеріорних ймовірностей до апріорних [11, 12]. Перевага однієї моделі над іншою визначається у відповідності із значенням байєсівського коефіцієнта $BF(i,j)$. Якщо даний коефіцієнт суттєво більше 1, то приймається рішення про прийняття або відхилення моделі. Для вибору моделі використовують критерій найбільшої граничної щільності розподілу $p(x|M_i)$, що відповідає умові $BF(i,j) > 1$.

Коректне застосування методики Бокса-Дженкінса забезпечує побудову адекватної математичної моделі процесу, якщо експериментальні дані відповідають *вимогам представництва* та *інформативності*. Перша вимога означає, що вибірка даних повинна охоплювати досить довгий проміжок часу, щоб повністю відображати поведінку того режиму функціонування процесу, для яких будується модель. Вимога інформативності означає, що вибірка повинна містити в собі об'єм інформації, достатній для оцінювання коефіцієнтів моделі. Наприклад, якщо моделюється процес другого порядку, то вибірка повинна забезпечувати коректне обчислення першої та другої похідної. Іноді інформативність формально оцінюють за допомогою величини дисперсії процесу, а також за кількістю гармонічних

складових, які містяться у процесі. Чим більше гармонік містить вибірка, тим вищою є її інформативність.

5. Результати досліджень

Для дослідження в якості експериментальних даних було застосовано статистичний ряд, структуру якого представлено в табл. 1.

Для вихідних даних, представлених в табл. 1, припустимо, що залежна змінна є нормально розподіленою, а в якості функції зв'язку візьмемо log-функцію.

Гістограма відповідності залежної змінної «*Збитки*» нормальному закону розподілу представлена на рис. 1. Дослідження проведено за допомогою пакету обробки статистичних даних *Eviews*.

Результат оцінювання лог-нормальної моделі представлено на рис. 2.

Середнє значення залежної змінної «Збитки» становить 1877,531. Інформаційний критерій Акайке слугує для вибору найкращої моделі із певного набору альтернативних моделей. Чим *менше* значення інформаційного критерію Акайке, тим краще побудовано модель відповідно до вихідних даних. Для лог-нормальної моделі інформаційний критерій Акайке становить 20,666.

Таблиця 1

Структура статистичних даних

№	Характеристика даних	Кількісне значення
1	Загальний розмір вибірки	9546 точок
2	Залежна змінна	Розмір виплаченої страховки, тобто збитки
3	Регіон продажу полісу	Київ, АР Крим, Одеса
4	Рік випуску автомобіля	починаючи з 2006 року
5	Марка автомобіля	Mitsubishi, Toyota, ВАЗ

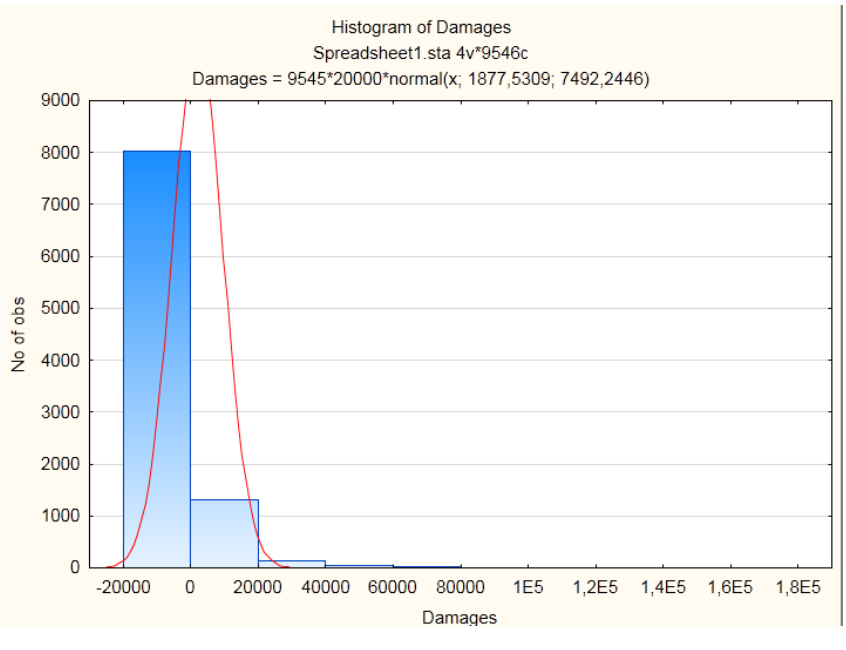

Рис. 1. Гістограма з нормально розподіленою залежною змінною

Dependent Variable: DAMAGES
Method: Generalized Linear Model (Quadratic Hill Climbing)
Date: 02/02/13 Time: 19:12
Sample: 1 9545
Included observations: 9545
Family: Normal
Link: Log
Dispersion computed using Pearson Chi-Square
Coefficient covariance computed using observed Hessian
Convergence achieved after 13 iterations

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-394.4486	118.1124	-3.339602	0.0008
AGE_OF_CAR	0.201732	0.058817	3.429811	0.0006
BRAND	-0.537131	0.056914	-9.437611	0.0000
LOCATION	-0.183465	0.049170	-3.731251	0.0002

Mean dependent var	1877.531	S.D. dependent var	7492.245
Sum squared resid	5.27E+11	Log likelihood	-98617.96
Akaike info criterion	20.66463	Schwarz criterion	20.66763
Hannan-Quinn criter.	20.66565	Deviance	5.27E+11
Deviance statistic	55192417	Restr. deviance	5.36E+11
LR statistic	165.7739	Prob(LR statistic)	0.000000
Pearson SSR	5.27E+11	Pearson statistic	55192417
Dispersion	55192417		

Рис. 2. Результат оцінювання лог-нормальної моделі

Гарною альтернативою критерію Акайке є критерій Хеннана-Куінна, якому притаманна швидка збіжність до істинного значення. Так як, одним із недоліків критерію Акайке є той факт, що оцінка Акайке непереконлива та асимптотичнопереоцінює (завищує) істинне значення з ненульовою ймовірністю, то було запропоновано більш переконливий критерій, який оснований на мінімізації відповідної суми, а не величини.

Інформаційний критерій Шварца зазвичай вибирає кращу модель з числом параметрів, котре не перевищує кількість параметрів в моделі, яка була обрана за критерієм Акайке. Критерій Шварца є асимптотично доцільним переконливим, в той час як критерій Акайке зміщений в сторону вибору параметризованих моделей. Для даного прикладу значення критерію Акайке, Хеннана-Куінна та Шварца однакові з точністю до 2-го знаку, тому для подальшого аналізу альтернативних моделей доцільно взяти будь-яке із них.

Значення стандартного відхилення залежної змінної – «Збитки» становить 7492,245.

Тест відношення правдоподібності (*LR statistic – Likelihood ratio test*) використовується для перевірки обмежень на параметри статистичної моделі, оцінених на основі вибіркових даних. Якщо значення *LR*-статистики більше критичного значення розподілу χ-квадрат при заданому рівні значимості, то обмеження відкидаються, перевага віддається моделі без обмежень, інакше – моделі з обмеженнями.

Значення дисперсії показує настільки в середньому випадкова величина відхиляється від математичного очікування, але важливо те, що відображає не в звичайних одиницях, а в квадратних. Однак сама дисперсія не дуже зручна для практичного аналізу, оскільки вона має розмірність квадрату випадко-

вої величини. Даний недолік виправлено за допомогою величини стандартного відхилення, яке в подальшому використовуємо для визначення величини ризику.

З точки зору фінансового аналізу первинним є значення стандартне відхилення, суть якого полягає в наступному – середнє відхилення результатів продажу продажів полісу (або страхування) від очікуваної доходності продажів полісу (страхування), тобто по суті ризик страхування. В якості міри ризику можна використовувати і величину середнього абсолютного відхилення (*mad*). На практиці значення стандартного відхилення більше від середнього абсолютного відхилення, але це величина однакового порядку, має місце наступне співвідношення:

$$mad = 0.7979 \cdot S.$$

Результат прогнозування величини збитків та значення ризику представлено на рис. 3. Відносна похибка результатів прогнозування за допомогою лог-нормальної моделі становить 1,06 %, а це свідчить про те, що прогноз здійснено з високою точністю, то ж проаналізуємо сумарні значення та результати оцінки приближення даних даною моделлю.

Сумарне прогнозне значення збитків дорівнює 18111231.380, а реальне – 17921032.581, що говорить про те, що припущення до залежна змінна «Збитки» є нормально розподіленою, а функція зв'язку – логарифмічна не є зовсім доцільним, але припустимим для подальшого аналізу альтернативних моделей.

Отже, побудована лог-нормальна модель є допустимою, але не найкращою для такого набору статистичних даних, тому доречно продовжити пошук максимально наближеної до даних моделі.

Результати побудови математичних моделей із використанням запропонованої структури наведено в табл. 2.

Таблиця 2

Результати побудови математичних моделей

№ п/п	Характеристики математичної моделі		Сумарні прогнозні значення збитків	Реальні сумарні значення збитків	Показник відхилення експериментальних даних від величини прогнозу	Ризик втрат
	Характер розподілу початкової змінної	Функція зв'язку				
1	Гамма	LOG	102008320,905	17921032,581	84087288,32	1,301
2	Нормальний	LOG	18111231,380		190198,799	0,495
3	Пуассона	LOG	17921032,574		0,007	0,547
4	Нормальний	тотожна	17921032,589		0,009	0,532

Результати оцінювання побудованих математичних моделей із використанням класичного та байєсівського підходів представлено в табл. 3.

В результаті побудови математичних моделей за допомогою припущень про початковий розподіл залежної змінної та підбору функції зв'язку – як однієї із складових УЛМ отримано:

1) найкращим наближенням даних до побудованої моделі є модель з початковим розподілом Пуассона залежної змінної та логарифмічною функцією зв'язку, що підтверджується прогнозним значенням сумарних збитків – 17921032,574;

2) найбільш точний прогноз було отримано за допомогою моделі з розподілом Пуассона та логарифмічною функцією зв'язку про що свідчить нульове значення відносної похибки;

3) величина ризику для побудованих моделей в середньому коливалась від 40–60 %, яке є гранично допустимим та все ж вимагає проведення додаткових заходів щодо мінімізації даної величини;

4) модель з нормальним розподілом та логарифмічною функцією зв'язку має мінімальне значення ризику (49–50 %), але значні відхилення від реальних даних про що свідчать результати оцінки даної моделі;

5) при порівнянні моделей с нормальним розподілом та логарифмічною і тотожною функцією зв'язку було виявлено, що інформаційний критерій Акайке приймає приблизно однакове значення 20,66, тому вибирати модель доречно, виходячи із сумарного результату прогнозу величини збитків;

6) модель з гамма розподілом та логарифмічною функцією зв'язку має найбільші відхилення (на 84087288,324) від реальних даних та найбільше значення похибки (більше 100 %);

7) із табл. 3 помітно, що результати оцінок, отриманих за допомогою байєсівського підходу нормального розподілу «близькі» до класичних результатів, отриманих за допомогою

Рис. 3. Результат прогнозування за допомогою лог-нормальної моделі

методу максимальної правдоподібності, але з більш точними показниками дисперсії та стандартного відхилення; крім цього, коефіцієнт детермінації, отриманий в результаті оцінюванням за допомогою байєсівського підходу показує приблизно однакові результати без вагомих «стрибків».

Таблиця 3

Результати оцінювання моделей

№ п/п	Класичний підхід			Байєсівський підхід			
	Mean	Std. deviance	Vari-ance,%	Mean	Std. deviance	Vari-ance, %	R-squared
1	11805,69	15358,12	130,091	11804,346	15247,237	128,669	0,89735
2	1897,457	939,91	49,535	1897,294	939,94	49,4	0,99854
3	1877,531	1027,567	54,73	1877,301	1027,552	55,679	0,99887
4	1877,531	999,302	53,224	1876,909	999,751	53,809	1

Отже, адекватною, прийнятною для практичного використання, є модель із законом розподілу Пуассона та експоненціальною функцією зв'язку через мінімальну величину похибки, показники значущості моделі, максимально наближеним до реальних даних прогнозних значень, достовірну оцінку величини ризику з використанням байєсівського підходу вибору кращої моделі та оцінювання невідомих параметрів УЛМ.

6. Висновки

Виконано дослідження щодо пошуку удосконаленої методики побудови математичних моделей для актуарних та фінансово-економічних процесів довільної природи. Запропоновано та експериментально доведено ефективність функціонування створеної багатокрокової методики із використанням математичного апарату УЛМ.

Розглянутий приклад ілюструє, що запропонована методика побудови математичних моделей є ефективним та зручним інструментом моделювання актуарних процесів. Для оцінювання невідомих параметрів УЛМ зручно використовувати байєсівський підхід, оперуючи апріорними та апостеріорними розподілами параметрів та алгоритмами вибору кращої моделі. Залучення новітніх комбінованих методів до оцінювання невідомих параметрів УЛМ та вибору кращої моделі на основі максимального наближення прогнозного значення до реальних даних розкриває нові горизонти досліджень властивостей сучасних методик та математичних методів. Надалі необхідно дослідити точність отриманих оцінок, їх збіжність, виконати порівняння з методами Монте-Карло для марковських ланцюгів та ін.

Застосування запропонованої моделі прогнозування процесів у сфері страхування за допомогою УЛМ та байєсівського підходу до оцінювання невідомих параметрів УЛМ гарантує високу точність оцінювання досліджуваної величини з мінімальним значенням фінансового ризику. Також можна зробити висновок, що сфера страхування, за умови належного менеджменту із застосуванням сучасних математичних методів, може бути надійним джерелом стабілізації економіки країни.

Література

1. Bowers, N. L. Actuarial mathematics [Text] / N. L. Bowers, H. U. Gerber, J. C. Hickman, D. A. Jones, C. J. Nesbitt. – Itasca (Illinois): Society of Actuaries, 1986. – 624 p.
2. Олексюк, О. С. Системи підтримки прийняття фінансових рішень на макрорівні [Текст] / О. С. Олексюк. – К.: Наукова думка, 1998. – 508 с.
3. Бідюк, П. І. Проектування комп'ютерних систем підтримки прийняття рішень [Текст] / П. І. Бідюк, О. П. Гожий, Л. О. Коршевнюк. – Миколаїв: Чорноморський державний університет ім. Петра Могили, 2011. – 320 с.
4. Gill, J. Generalized linear models: a unified approach [Text] / J. Gill. – USA: New Delli, 2001. – 110 p.
5. Бідюк, П. І. Оцінювання параметрів моделей із застосуванням методу Монте-Карло для марковських ланцюгів [Текст] / П. І. Бідюк, А. С. Борисевич, // Наукові праці Миколаївського державного гуманітарного університету ім. Петра Могили. 2008. – № 77. – С. 21–37.
6. Tsay, S. Financial time series analysis [Text] / S. Tsay. – Hoboken (New Jersey): John Wiley & Sons, 2010. – 715 p.
7. Enders, W. Applied econometric time series [Text] / W. Enders. – New York: Wiley and Sons, 1994. – 433 p.
8. Бідюк, П. І. Аналіз часових рядів [Текст] : навч, посібник / П. І. Бідюк, В. Д. Романенко, О. Л. Тимощук. – К: НТУУ «КПІ», 2013. – С. 115–158.
9. McCullagh, P. Generalized Linear Models [Text] / P. McCullagh, J. A. Nelder. – New York: Chapman & Hall, 1990. – 526 p. doi: 10.1007/978-1-4899-3242-6
10. Трухан, С. В. Прогнозування актуарних процесів за допомогою узагальнених лінійних моделей [Текст] / С. В. Трухан, П. І. Бідюк // Наукові вісті НТУУ «КПІ». – 2014. – № 2. – С. 14–20.
11. Bergman, N. Recursive Bayesian Estimation: Navigation and Tracking Applications [Text] / N. Bergman // Linkoping University (Sweden). – 1999. – Vol. 579. – P. 219.
12. Besag, J. Markov Chain Monte Carlo for Statistical Inference [Text] / J. Besag // Working Paper, Center for Statistics and the Social Sciences. – 2001. – Vol. 9. – P. 25.